Západočeská univerzita v Plzni
Fakulta aplikovaných věd
Katedra informatiky a výpočetní techniky

Semestrální práce z předmětů KIV/AZS a KIV/TKS

PLP parametrizace pro ASR systém JLASER

Plzeň, 2009

Tomáš Brychcín
A08N0047P
brychcin@students.zcu.cz

Obsah

1 Úvod

Hlavním cílem této práce byla implementace PLP (Perceptual Linear Predictive) parametrizace do ASR (Automatic Speech Recognition) systému JLASER vyvíjeného Laboratoří Inteligentních Komunikačních Systémů (LIKS) na Západočeské Univerzitě v Plzni.

Sytém JLASER poskytuje sadu nástrojů pro automatické ropoznávání řeči a je zpracovaný v jazyce JAVA. Rozpoznávač JLASER je založen na hybridní architektuře kombinující výhody umělé neuronové sítě a skrytých Markovových modelů. Proces rozpoznávání tímto systémem lze rozdělit do třech částí. První je zpracování vstupního řečového signálu, neboli parametrizace. Vstupní data, vzniklé navzorkováním řečového signálu, mohou obsahovat kromě čisté promluvy i různý šum na pozadí. Úkolem parametrizace je ze vstupních dat extrahovat informaci popisující pouze promluvu a to pomocí omezeného množství hodnot. Výstupem parametrizace je posloupnost příznakových vektorů. Každý příznakový vektor charakterizuje úsek nahrávky o délce přibližně 16ms. Dále se pomocí klasifikátoru (neuronové sítě) počítá pravděpodobnost s jakou aktuální příznakový vektor představuje konkrétní foném. Dekodér poté na základě těchto pravděpodobností rozhodne co bylo řečeno. Celková úspešnost rozpoznávání tedy závisí jak na kvalitě klasifikátoru, tak na kvalitě parametrizace.

Literatura uvádí, že PLP parametrizace by měla dávat lepší výsledky na zašuměném signálu než MFCC parametrizace, která je již v systému JLASER implementována.

2 PLP parametrizace

PLP (Perceptual Linear Predictive) je technika pro extrakci příznaků z řečového signálu. PLP se snaží přizpůsobit zpracování signálu způsobu vnímání lidského sluchu. Zohledňují se zde 3 základní faktory z psychofyziky slyšení (viz [1]): kritické pásmo spektrální citlivosti, křivky stejné hlasitost a vztah mezi intenzitou a vnímanou hlasitostí. Výsledkem jsou koeficienty popisující vyhlazený tvar spektra řečového signálu.

2.1 Lineární predikce

Princip lineární predikce spočívá v tom, že se pokusíme odhadnout n-tý vzorek signálu $x(n)$ pomocí lineární kombinace Q předchozích prvků tohoto

signálu

$$\hat{x}\,(n) = \sum_{k=1}^{Q} -a\,(k)\,x\,(n-k). \tag{1}$$

Při zpracování řečových signálů se nejčastěji využívá autokorelační přístup, který předpokládá, že signál je nulový vně zkoumaného segmentu. Při odhadování signálu vzniká tzv. chyba predikce. Funkce krátkodobé energie chyby signálu je definována jako

$$E = \sum_{n} \left[x\,(n) - \hat{x}\,(n)\right]^2 = \sum_{n} \left[x\,(n) - \sum_{k=1}^{Q} -a\,(k)\,x\,(n-k)\right]^2. \tag{2}$$

Pokud zavedeme že koeficient $a\,(0) = 1$, budeme moci enerigii chyby signálu vyjádřit jako

$$E = \sum_{n} \left[\sum_{k=0}^{Q} a\,(k)\,x\,(n-k)\right]^2. \tag{3}$$

Pro co nejlepší výsledky je nutné minimalizovat tuto energii chyby E a to tak, že položíme parciální derivace podle všech koeficientů $a\,(i)$ rovné 0

$$\frac{\partial E}{\partial a\,(i)} = 0, \qquad 1 \leq i \leq Q. \tag{4}$$

Řešením této rovnice dostaneme soustavu rovnic

$$\frac{\partial E}{\partial a\,(i)} = 0 = \sum_{k=0}^{Q} a\,(k) \sum_{n} x\,(n-k)\,x\,(n-i), \qquad 1 \leq i \leq Q, \tag{5}$$

což lze vyjádřit pomocí autokorelačních funkcí $r\,(i)$

$$0 = \sum_{k=0}^{Q} a\,(k) r\,(k-i)\,, \qquad 1 \leq i \leq Q, \tag{6}$$

kde

$$r\,(i) = \sum_{n} x\,(n) x\,(n-i)\,. \tag{7}$$

Ze vztahu 2 lze odvodit rovnici pro energii chyby predikce

$$E = \sum_{k=0}^{Q} a\,(k)\, r\,(k). \tag{8}$$

Ze vztahů 6 a 8 dostáváme soustavu rovnic, kterou lze zapsat v maticovém tvaru jako

$$\begin{bmatrix} r\,(0) & r\,(1) & r\,(2) & \cdots & r\,(Q) \\ r\,(1) & r\,(0) & r\,(1) & \cdots & r\,(Q-1) \\ r\,(2) & r\,(1) & r\,(0) & \cdots & r\,(Q-2) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ r\,(Q) & r\,(Q-1) & r\,(Q-2) & \cdots & r\,(0) \end{bmatrix} \begin{bmatrix} 1 \\ a\,(1) \\ a\,(2) \\ \vdots \\ a\,(Q) \end{bmatrix} = \begin{bmatrix} E \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}. \tag{9}$$

Řešením této soustavy rovnic obdržíme potřebné koeficienty $a\,(k)$ lineární predikce. Matice autokorelačních koeficientů je v Töplitzově tvaru. Pro řešení je možné použít například Levinsonův-Durbinův algoritmus popsaný v následující podkapitole 2.1.1.

2.1.1 Levinsonův-Durbinův algoritmus

Levinsonův-Durbinův algoritmus je vysoce efektivní iterativní proces pro řešení soustavy rovnic ve tvaru

$$R\vec{a} = \vec{b}, \tag{10}$$

kde R je matice v Töplitzově tvaru (čtvercová symetrická matice, se stejnými hodnotami na všech diagonálách ve směru hlavní diagonály), $\vec{b}$ je známý vektor a $\vec{a}$ je vektor, který hledáme.

Definujeme reverzní vektor $v^{\#}$ k vektoru v jako vektor, kde prohodíme první prvek s posledním, druhý s předposledním atd. Matice R v Töplitzově tvaru má následující vlastnost

$$R\vec{a} = \vec{b} \Rightarrow R\vec{a}^{\#} = \vec{b}^{\#}. \tag{11}$$

Dále superscript p bude značit, že se jedná o p iteraci. Označme čtvercovou matici o velikosti $p \times p$, která kopíruje levý horní blok matice R, jako R^p. Matice R^p je opět v Töplitzově tvaru. Vektor odhadů řešení v p iteraci o velikosti

p označíme jako $\vec{a}^p$. Předpokládejme, že $\vec{a}^p = \begin{bmatrix} 1 & a^p(1) & a^p(2) & \cdots & a^p(p-1) \end{bmatrix}^T$ a že $R^p\vec{a}^p = \vec{b}^p$, kde $\vec{b}^p = \begin{bmatrix} E^p & 0 & 0 & \cdots & 0 \end{bmatrix}^T$. Neboli

$$\begin{bmatrix} r(0) & r(1) & \cdots & r(p-1) \\ r(1) & r(0) & \cdots & r(p-2) \\ \vdots & \vdots & \ddots & \vdots \\ r(p-1) & r(p-2) & \cdots & r(0) \end{bmatrix} \begin{bmatrix} 1 \\ a^p(1) \\ \vdots \\ a^p(p-1) \end{bmatrix} = \begin{bmatrix} E^p \\ 0 \\ \vdots \\ 0 \end{bmatrix}. \quad (12)$$

V iteraci $p+1$ rozšíříme matici R^p na R^{p+1} a do vektoru odhadů řešení $\vec{a}^p$ přidáme 0. Bude platit

$$\begin{bmatrix} r(0) & r(1) & \cdots & r(p-1) & r(p) \\ r(1) & r(0) & \cdots & r(p-2) & r(p-1) \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ r(p-1) & r(p-2) & \cdots & r(0) & r(1) \\ r(p) & r(p-1) & \cdots & r(1) & r(0) \end{bmatrix} \begin{bmatrix} 1 \\ a^p(1) \\ \vdots \\ a^p(p-1) \\ 0 \end{bmatrix} = \begin{bmatrix} E^p \\ 0 \\ \vdots \\ 0 \\ q^{p+1} \end{bmatrix}, \quad (13)$$

kde $q^{p+1} = \sum_{i=0}^{p} a^p(i) \cdot r(p-i)$. Z principu linearity a z vlastností matice v Töplitzově tvaru plyne

$$R^{p+1}\left(\begin{bmatrix} 1 \\ a^p(1) \\ \vdots \\ a^p(p-1) \\ 0 \end{bmatrix} + k^{p+1}\begin{bmatrix} 0 \\ a^p(p-1) \\ \vdots \\ a^p(1) \\ 1 \end{bmatrix}\right) = \left(\begin{bmatrix} E^p \\ 0 \\ \vdots \\ 0 \\ q^{p+1} \end{bmatrix} + k^{p+1}\begin{bmatrix} q^{p+1} \\ 0 \\ \vdots \\ 0 \\ E^p \end{bmatrix}\right). \quad (14)$$

Konstantu k^{p+1} volíme tak, aby $q^{p+1} + k^{p+1}E^p = 0$, z čehož plynou následující vztahy

$$k^{p+1} = -\frac{1}{E^p}\sum_{i=0}^{p} a^p(i) \cdot r(p-i), \quad (15)$$

$$E^{p+1} = E^p\left[1 - \left(k^{p+1}\right)^2\right], \quad (16)$$

$$a^{p+1}(i) = a^p(i) + k^{p+1} \cdot a^p(p-i), \qquad 0 \le i \le p. \quad (17)$$

V první iteraci se hodnoty volí následovně, $E^1 = r(0)$ a $a^1(0) = 1$.

2.2 Segmentace a váhování signálu

Před parametrizací je signál nejdříve rozdělen do tzv. rámců o délce N vzorků. Současně je použito 50% překrývání rámců. Následně se provádí váhování signálu, které potlačí náhlé uříznutí vzorků signálu. Je použito Hammingovo okénko dané vztahem

$$w[n] = \begin{cases} 0.54 - 0.46\cos\frac{2\pi n}{N-1} & \text{pro } 0 \leq n \leq N-1 \\ 0 & \text{jinde} \end{cases} \tag{18}$$

kde N je délka rámce. Vzorky v rámci jsou váženy podle následujícího vztahu

$$s[n] = x[n] \cdot w[n] \qquad 0 \leq n \leq N-1. \tag{19}$$

2.3 Výpočet výkonového spektra

Poté co máme singál rozdělený na rámce, spočítáme pro každý rámec jeho výkonové spektrum pomocí diskrétní Fourierovy transformace (DFT). DFT je počítána podle algoritmu rychlé Fourierovy transformace (FFT).

$$P(f) = |DFT\{s[n]\}|^2 = [\mathrm{DFT}_{\mathrm{Re}}(s[n])]^2 + [\mathrm{DFT}_{\mathrm{Im}}(s[n])]^2 \qquad 0 \leq n \leq N-1, \tag{20}$$

kde N je délka rámce.

2.4 Průchod kritickými pásmovými filtry

2.4.1 Frekvenční maskování zvuků

Frekvenční rozsah zvuku, který většina lidí vnímá, začíná kolem 16 Hz a dosahuje ke 20 kHz. Schopnost odlišit dva frekvenčně blízké tóny je ovlivněna tzv. frekvenčním maskováním. Pokud znějí dva tóny současně, může jeden z nich potlačit (maskovat) slyšitelnost toho druhého. Úroveň maskovaní je závislá na frekvenční vzdálenosti obou signálů. Šířka pásma, ve kterém je daný zvuk maskován se nazývá šířka kritického pásma. Podrobněji v [2]. Schopnost ucha rozlišovat tóny směrem k vyšším frekvencím klesá přibližně logaritmicky. To má za následek, že šířka kritického pásma se mění v závislosti na frekvenci. Kritické pásmo má na nejnižších kmitočtech velikost kolem 100 Hz, zatímco na nejvyšších kmitočtech dosahuje až 4 kHz.

Simulování těchto jevů je v PLP realizováno transformací frekvenční osy $f[Hz]$ do Barkova měřítka $\Omega[bark]$ podle vztahu

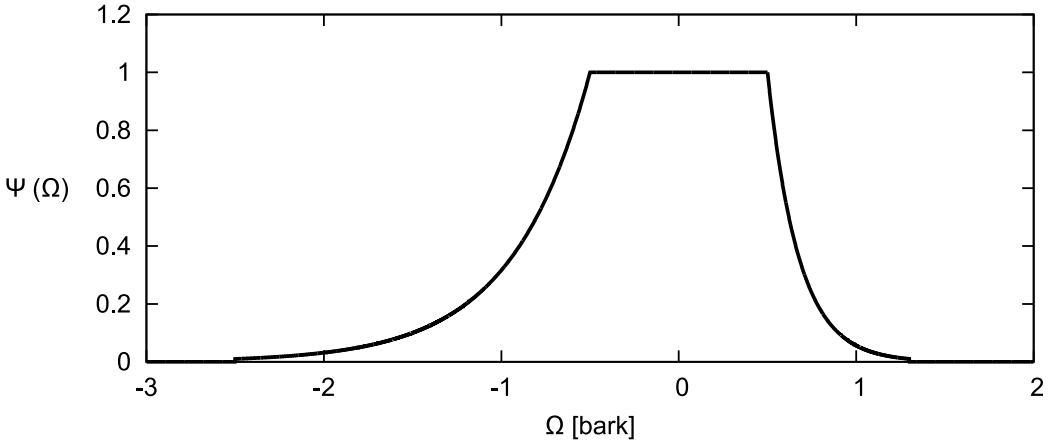
$$\Omega\left(f\right) = 6\ln\left(\frac{f}{600} + \sqrt{\left(\frac{f}{600}\right)^2 + 1}\right). \tag{21}$$

Pro převod z Barkova měřítka zpět do frekvenční osy je možné použít inverzní vztah

$$\Gamma\left(\Omega\right) = 300\left(e^{\frac{\Omega}{6}} - e^{-\frac{\Omega}{6}}\right). \tag{22}$$

Dále jsou vytvořeny maskující křivky simulující kritická pásma slyšení. Maskující křivku (viz obrázek 1) lze podle [1] popsat vztahem

$$\Psi\left(\Omega\right) = \begin{cases} 0 & \text{pro } \Omega < -2.5 \\ 10^{\Omega+0.5} & \text{pro } -2.5 \leq \Omega \leq -0.5 \\ 1 & \text{pro } -0.5 < \Omega < 0.5 \\ 10^{-2.5(\Omega-0.5)} & \text{pro } 0.5 \leq \Omega \leq 1.3 \\ 0 & \text{pro } \Omega > 1.3 \end{cases} \tag{23}$$



Obrázek 1: Maskující křivka simulující kritické pásmo slyšení

Tyto křivky jsou na frekvenční ose v Barkově měřírku rozmístěny lineárně s krojem přibližně 1 bark. První a poslední filtr jsou umístěny tak, aby měly střed v mezních hodnotách frekvence přenášeného pásma (viz obrázek 2). Doporučené hodnoty pro rozmístění těchto filtrů jsou uvedeny v tabulce 1.

Tabulka 1: Doporučené hodnoty rozmístění filtrů (maskovacích křivek) pro konkrétní šířku pásma (podle [1])

Vzorkovací frekvence F_v [kHz]	Přenášené pásmo (0 až B_w) [kHz]	Přenášené pásmo (0 až B_{bw}) [bark]	Počet filtrů M	Krok rozmístění filtrů [bark]
8	0 - 4	0 - 15.57	15 + 2	973
11	0 - 5.5	0 - 17.47	17 + 2	971
16	0 - 8	0 - 19.71	19 + 2	985
22	0 - 11	0 - 21.62	21 + 2	983
44	0 - 22	0 - 25.77	25 + 2	991

2.4.2 Aproximace křivky stejné hlasitosti

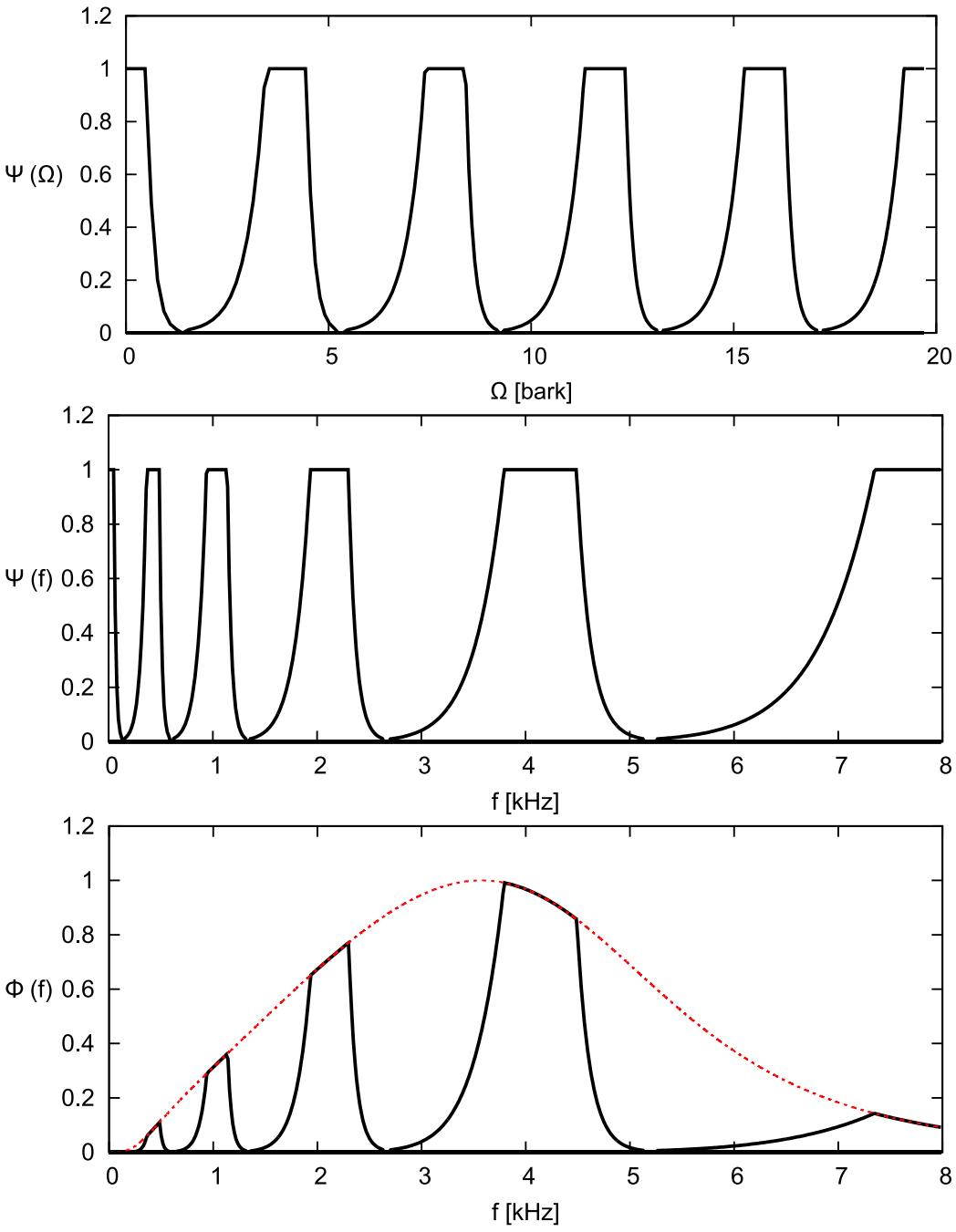
Člověk vnímá intenzitu zvuku v závislosti na frekvenci jako hlasitost zvuku. Hlasitost zvuku je zcela subjektivní pocit, kterým člověk posuzuje intenzitu zvuku. Tento jev je popsán tzv. křivkami stejné hlasitosti, které udávají, jaká intenzita způsobí na konkrétních frekvencích stejný vjem hlasitosti. Další krok PLP analýzy je aproximace těchto křivek pro hladinu hlasitosti 40Ph. Tato aproximace je dána vztahem

$$E(\omega) = \frac{\omega^4 \left(\omega^2 + 56.9 \cdot 10^6\right)}{\left(\omega^2 + 6.3 \cdot 10^6\right)^2 \left(\omega^2 + 379.4 \cdot 10^6\right) \left(\omega^6 + 9,6 \cdot 10^{26}\right)}, \qquad (24)$$

kde $\omega = 2\pi f$. Lidský sluch je nejcitlivější na zvuky o frekvenci 3-4 kHz, proto maximum této funkce je přibližně na frekvenci $f \cong 3600Hz$. Hodnoty pásmových filtrů jsou násobeny hodnotami aproximující křivky podle vzorce

$$\Phi_m(f) = E(2\pi f) \cdot \Psi\left(\Omega(f) - \Omega_m\right), \qquad 1 \leq m \leq M, \qquad (25)$$

kde Ω_m je střed m-tého kritikého pásmového filtru. Průběhy kritických pásmových filtrů jsou vykresleny na obrázku 2.



Obrázek 2: Rozložení kritických pásmových filtrů

Šířka pásma je $B_w = 8$kHz ($B_{bw} = 19{,}71$ bark) při vzorkovací frekvenci $F_v = 16$kHz. Na obrázcích je vyznačen každý čtvrtý filtr, aby se filtry vzájemně nepřekrývaly.

2.4.3 Sumarizace vzorků výkonového spektra

Průchod výkonového spektra $P(f)$ získaného Fourierovo transformací (viz. kapitola 2.3) m kritickým pásmovým filtrem Φ_m vyjádříme vztahem

$$\Xi_m = \sum_{f=f_{md}}^{f_{mh}} P(f)\,\Phi_m(f) \qquad 1 \leq m \leq M, \tag{26}$$

kde f_{md} a f_{mh} značí dolní a horní hranici m-tého kritického filtru. Tyto hodnoty lze určit z rovnice 22 jako $f_{md} = \Gamma(\Omega_m - 2.5)$ a $f_{mh} = \Gamma(\Omega_m + 1.3)$

2.4.4 Závislost hlasitosti na intenzitě zvuku

Hlasitost zvuku je úměrná intenzitě tohoto zvuku umocněné na 0.3 Pro respektování této vlastnosti provedeme třetí odmocninu na hodnotách získaných z jednotlivých kritických pásmových filtrů

$$\xi_m = (\Xi_m)^{0.3} \qquad 1 \leq m \leq M. \tag{27}$$

2.5 Výpočet autokorelačních koeficientů

Pro výpočet Q autokorelačních koeficientů z hodnot ξ_m použijeme následující rovnici

$$\begin{gathered} R(i) = \tfrac{1}{2(M-1)} \left\{ \alpha(i,0) + 2\left[\sum_{m=1}^{M-2} \alpha(i,m)\right] + \alpha(i,M-1) \right\}, \\ 0 \leq i \leq Q-1, \qquad 0 \leq m \leq M-1, \end{gathered} \tag{28}$$

kde $\alpha(i,m)$ je vypočítáno podle

$$\alpha(i,m) = \xi_m \cos\left(\frac{i \cdot m \cdot \pi}{M-1}\right). \tag{29}$$

2.6 Výpočet kepstrálních koeficientů

Z autokorelačních koefientů $R(i)$, spočítaných v předchozí kapitole jsou nejdříve vypočteny autoregresní koeficienty $a(i)$ pomocí Levinson-Durbinova algoritmu popsaného v kapitole 2.1.1. Tyto koeficienty popisují vyhlazený tvar spektra signálu.

Tyto koeficienty jsou následně převedeny na kepstrální koeficienty $c(z)$ pomocí následujících vztahů

$$\begin{aligned} c(0) &= E, \qquad \text{E je chyba lineární predikce} \\ c(1) &= -a(1), \\ c(k) &= -a(k) - \sum_{i=1}^{k-1} \tfrac{i}{k} c(i)\, a(k-i), \qquad 2 \le k \le Q, \\ c(k) &= -\sum_{i=1}^{Q} \tfrac{k-i}{k} c(k-i)\, a(i), \qquad k > Q. \end{aligned} \tag{30}$$

Důvodem k tomuto kroku jsou jejich rozdílné vlastnosti. Oboje koeficienty (autoregresní i kepstrální) popisují vyhlazený tvar spektra signálu, ale kepstrální koeficienty, na rozdíl od autoregresních koeficientů, jsou vzájemně velmi málo korelované, což je výhodné pro algoritmy klasifikace. Kepstrální koeficienty jsou koeficienty Fourierova rozvoje logaritmu amplitudového spektra a pro úcely rozpoznávání jsou robustnejší a vhodnejší. Kepstrální koeficienty jsou výsledkem PLP parametrizace aktuálního řečového rámce.

3 Výsledky prvních testů

Byly provedeny první testy, kde se u MFCC i PLP hledalo 13 příznaků. U PLP byly dále nastaveny tyto hodnoty:

- Počet kepstrálních koeficientů $K = 13$.
- Počet autoregresních koeficientů $Q = 15$.
- Počet kritických pásmových filtrů $M = 21$.

Úspěšnost rozpoznávání byla počítána podle vztahu

$$\%Corr = \frac{H}{N}, \tag{31}$$

kde H je počet správně rozpoznaných slov v aktuální promluvě a N je celkový počet slov.

Tabulka 2: Výsledky testů u MFCC a PLP parametrizace

	$\%Corr$	
Korpus	**MFCC**	**PLP**
šachy	97,69%	97,85%
vlaky	84,62%	84,01%

4 Závěr

PLP parametrizace byla do rozpoznávacího systému JLASER implementována a první testy dosahovaly přibližně stejné úspěšnosti jako u MFCC parametrizace (viz kapitola 3). Se systémem JLASER jsem se již seznámil v rámci několika semestrálních prací a hlavně v bakalářské práci a proto mi implementace nečinila větší problémy. Dalším nezbytným krokem pro kvalitní běh PLP parametrizace je stanovení počtu hledaných autoregresních a kepstrálních koeficientů, to už ale přesahuje tuto semestrální práci.

Reference

[1] Psutka, J. a Müller, L. a Matoušek, J. a Radová, V.: *Mluvíme s počítačem česky*, s. 752, Academia, Praha, 2006. ISBN 80-200-1309-1.

[2] Prchal, J. a Šimák, B.: *Digitální zpracování signálů v telekomunikacích*, Skripta ČVUT, Praha, 2000.